

DNA simulation, version 1

Paul Rakow

September 27, 2015

1 Introduction

This is an attempt to simulate the inheritance of DNA over several generations, to calculate how likely or unlikely matches of a given strength will be, for a given relationship.

The simulation is based on the classic laws of inheritance. This simulation only includes the 22 autosomal chromosome pairs (the X and Y chromosomes are not included). Cross-overs are assumed to take place randomly, with a probability of 1% per cM in each generation.

In this first version I have used the same cross-over rate for male and female parents — in the next version I will use a larger rate for females, and a smaller rate for males. I suspect this will not change results too much, especially for distant relatives, as the number of male links and female links will be roughly equal in most cases. (The number of distant cousins you have, with both of you descended along all-male or all-female lines, is very small compared with the total number of cousins you have.)

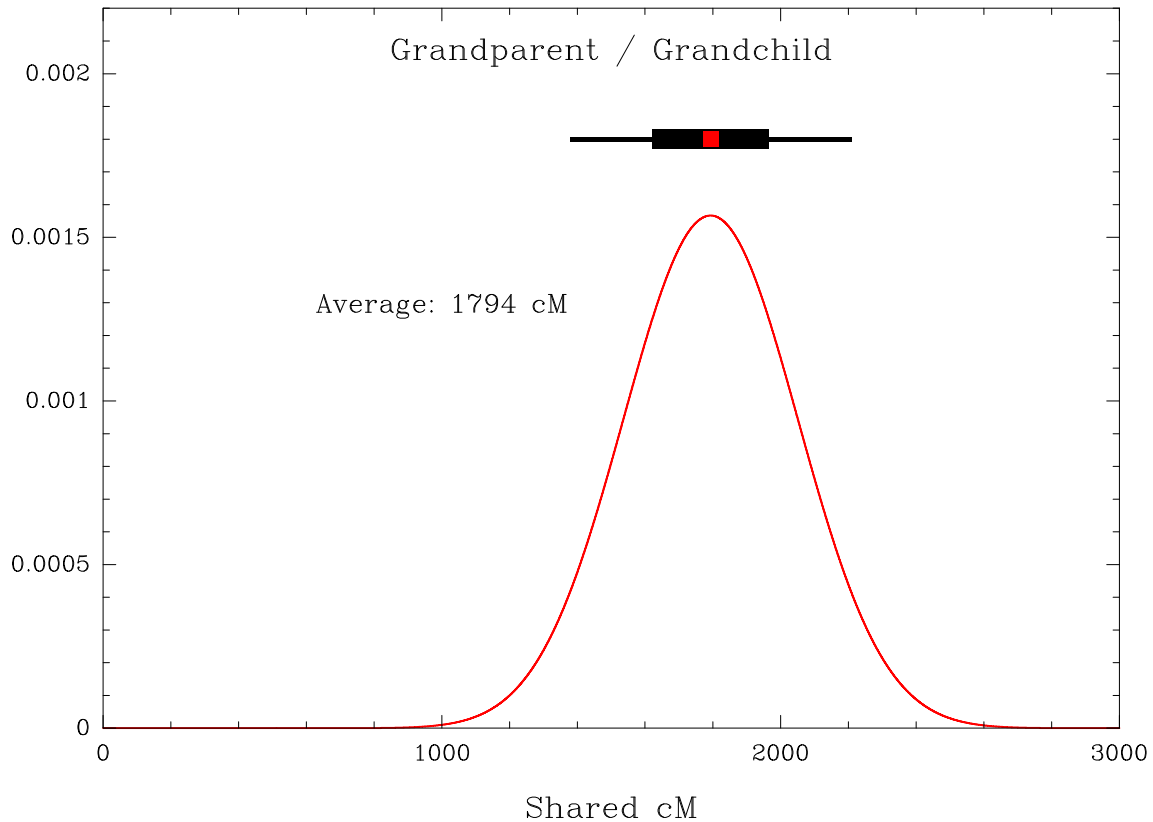
2 Matches between Ancestors and Descendants

When we compare one particular ancestor with a descendant, the average amount of shared DNA decreases by a factor of two each generation. In the first generation there is no random spread in this fraction, one chromosome of the pair comes from the mother, the other from the father. After this, randomness enters the process, the match with a grandparent will be 25% on average, but there can be a considerable random variation from this figure.

Knowing that a grandchild inherits on average 25% from each grandparent doesn't tell us whether this will consist mostly of entire unbroken chromosomes, or of many short segments. Calculating the number and size of the segments (and how large the spread around the averages will be) needs input from the expected rate of recombination.

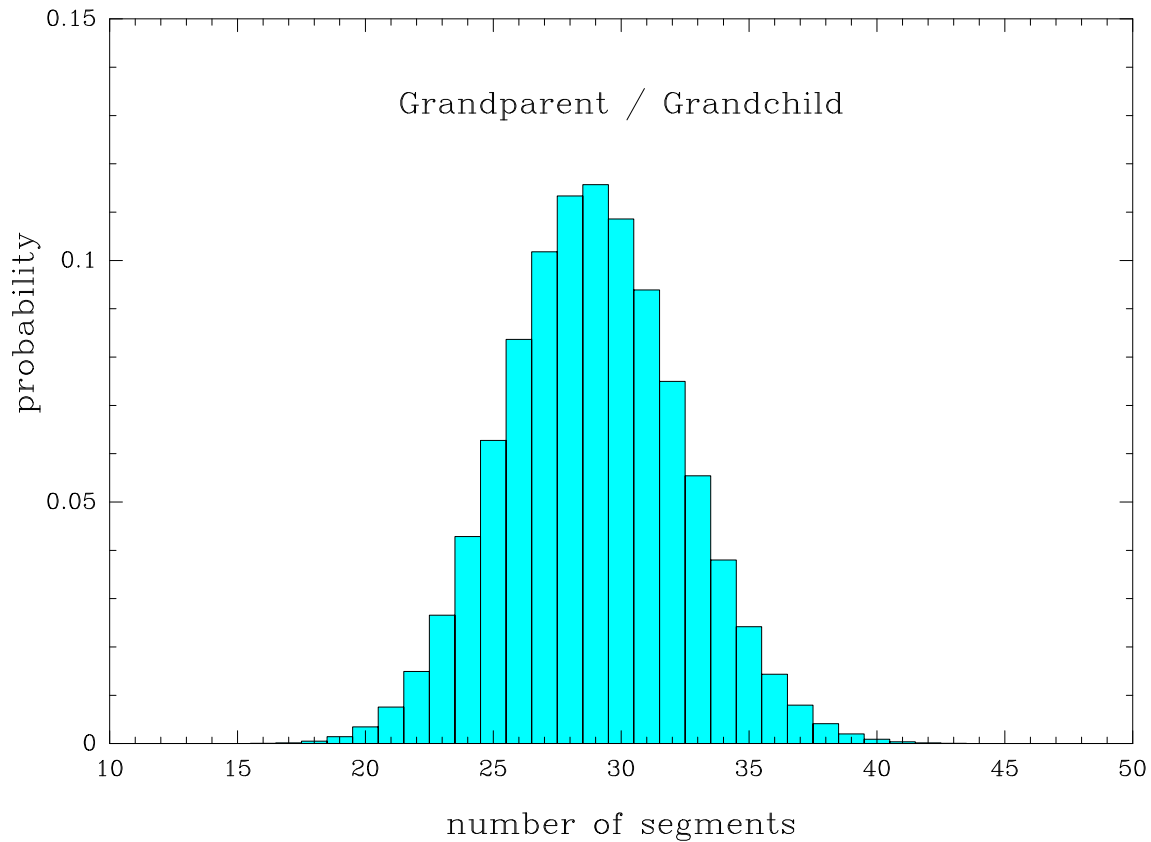
3 Grandparent/Grandchild

The simplest quantity to look at is the number of shared cM between a grandparent and grandchild. I simulated a million grandparent/grandchild pairs, and counted up how many of them had a given number of shared cM. A histogram of the results is shown below. The histogram is plotted with bins of 1 cM, we see a very smooth bell-shaped curve, as often happens when a large number of independent events are averaged over.



The bar above the histogram shows the spread of the results. The average is 1794 cM (the red spot near the centre of the bar). 50% of the grandchildren lie within the range of the thick part of the bar, from 1622 cM to 1965 cM, and 90% within the range of the thin part of the bar, 1377 cM to 2210 cM.

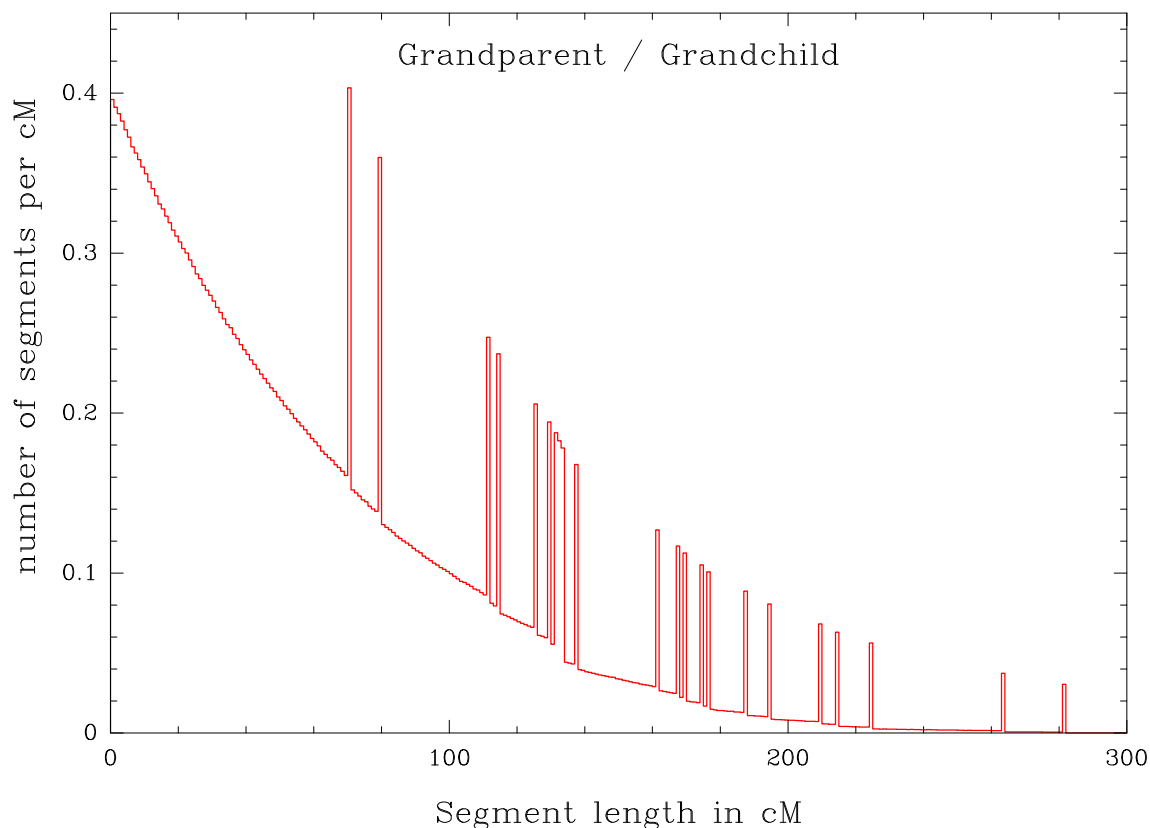
The next question is how many segments will a grandparent and grandchild share?



The histogram above shows the result, counting *all* matching (Identical By Descent) segments, however short. In that case, the average number of segments was 29. However, the numbers reported by DNA matching tools will normally have a threshold for reporting a segment. If we only count segments longer than 10 cM, the average segment number drops to 25.

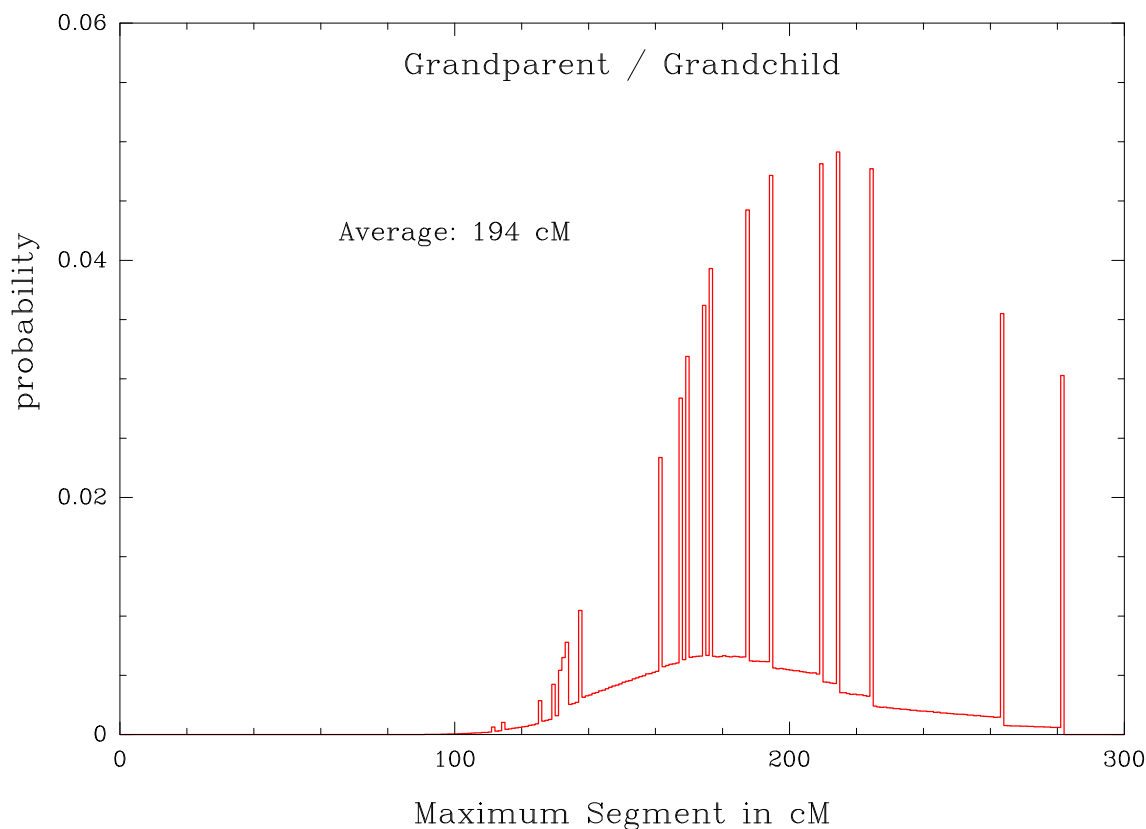
What would be the most useful threshold to use when counting segments?

How long are the segments, how many large segments, and how many small segments should we expect?



This time we see a very complicated looking graph. There is a fairly smooth background curve, dropping down, so that we see segments with lengths near 20 cM are about 3 times more common than segments with lengths near 100 cM. But we also see lots of spikes, where particular segment lengths are much more common than we would expect. These spikes are entire chromosomes that have been passed down as single segments, without crossover.

Finally, we take a look at the longest segment that a grandparent and grandchild will share.

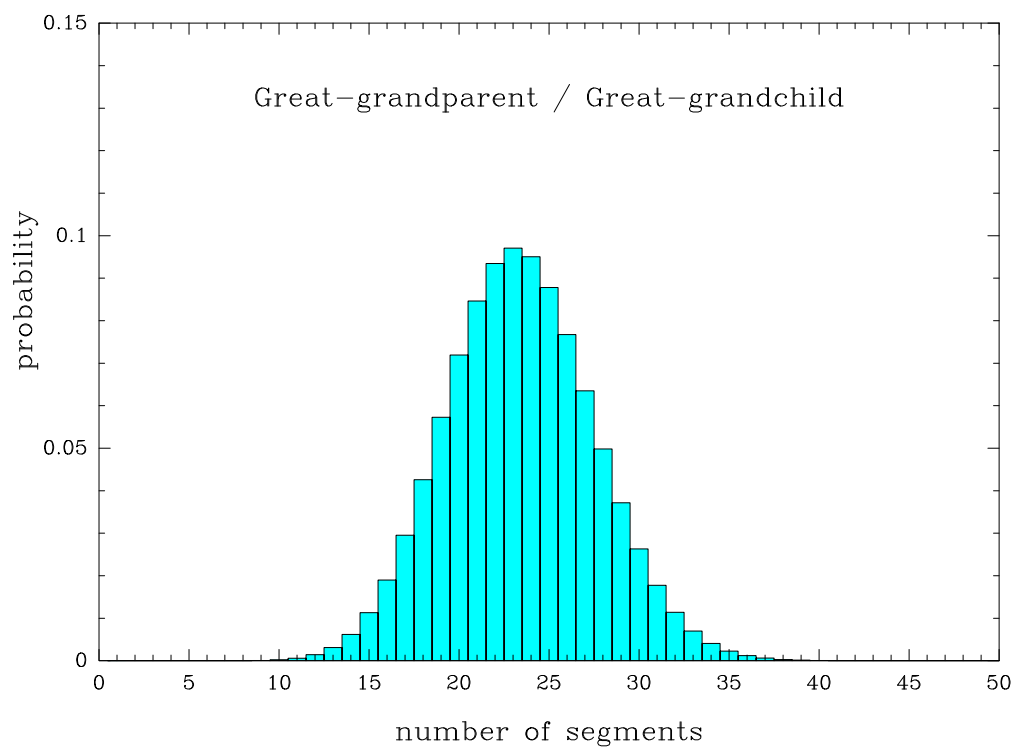
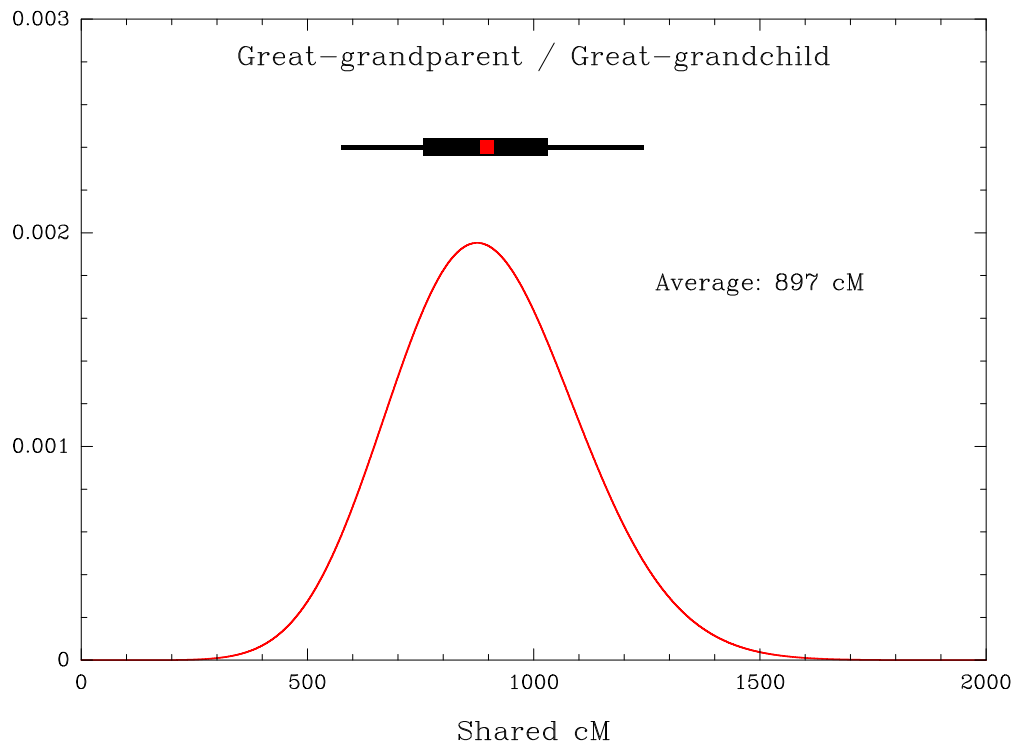


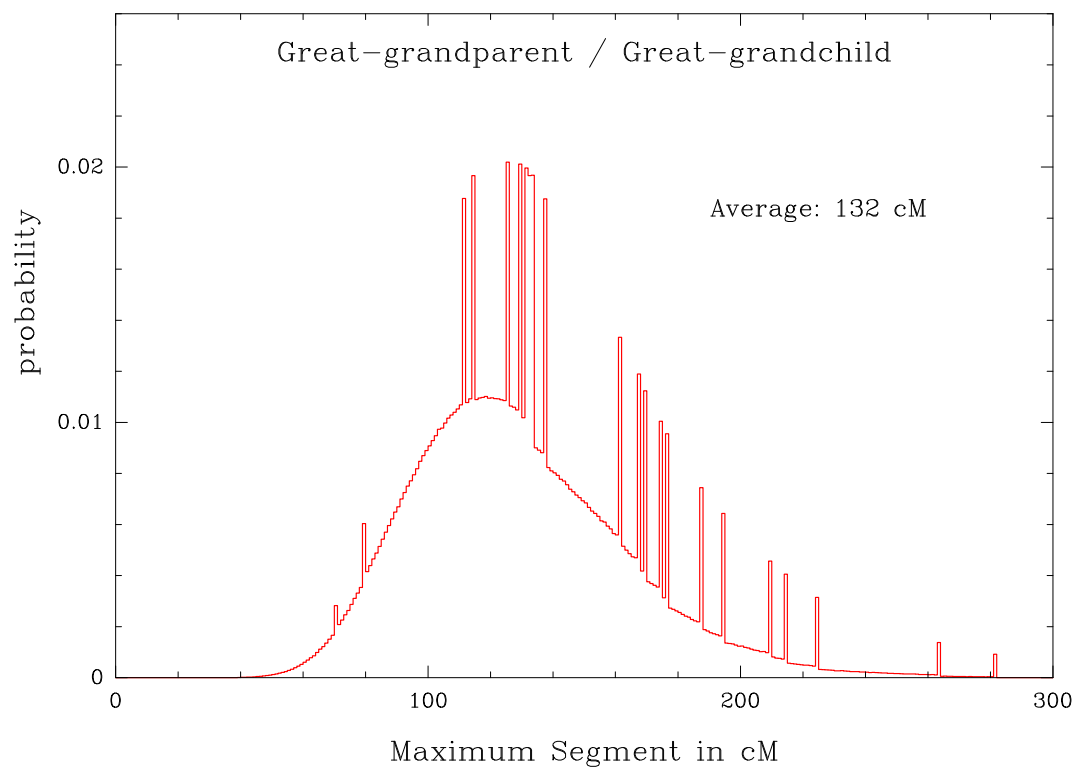
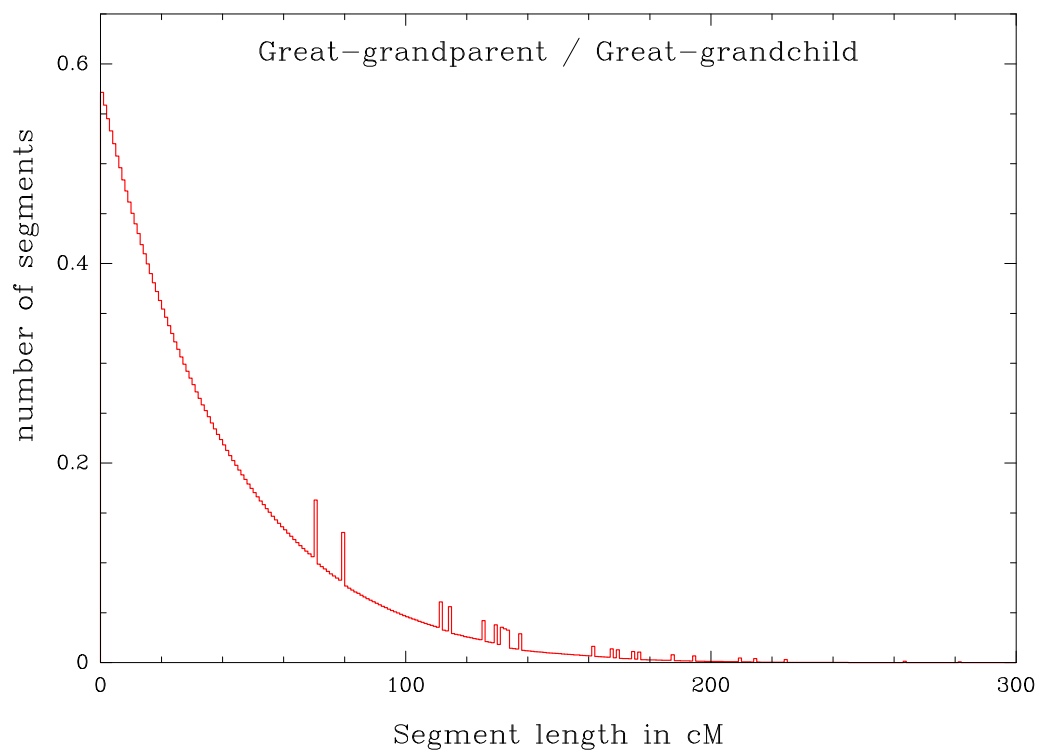
Again, this is a very complicated curve, with spikes representing cases where the longest segment is an entire chromosome, and a smoother curve in between, from cases where the longest segment is on a chromosome which has been split by a crossover.

There may be a problem with false negatives in the values for the longest segment. Sometimes the DNA tools will split a long segment into two, because there is a tiny region of mismatch somewhere along the chromosome. If this has happened, the number the tool returns for the longest segment might really be the length of the second or third longest segment. You can see this most easily in parent/child comparisons. The longest segment should always be chromosome 1, but quite often you will see a smaller number for the length of the longest segment.

4 Great-grandparent/Great-grandchild

I have made the same plots as before for the match between a great-grandparent and their great-grandchild.





5 Matches between Cousins

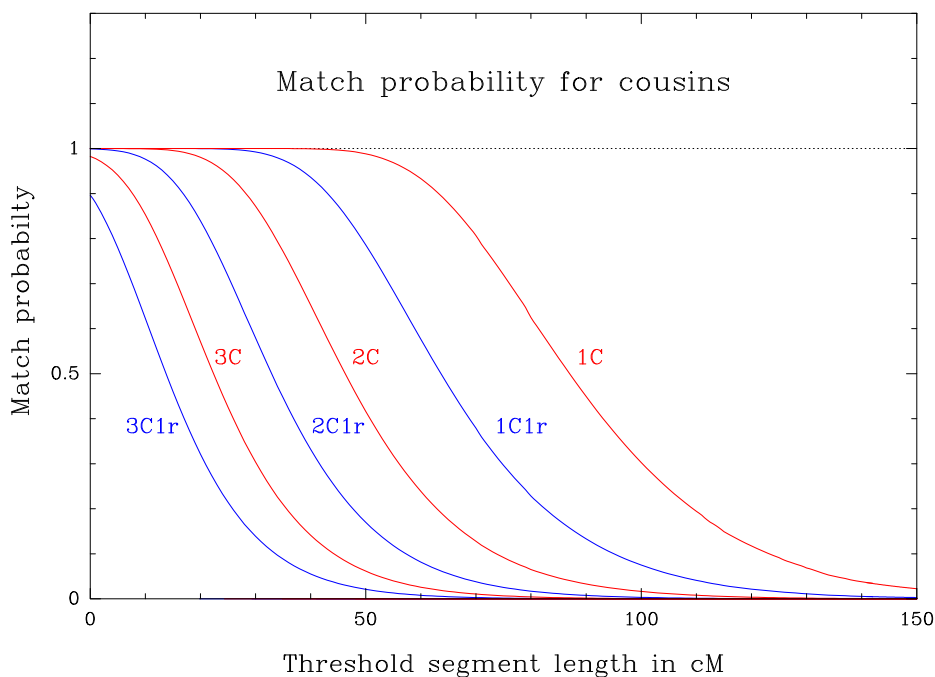
The main purpose of doing the simulation was to get an idea of how likely a distant cousin is to match with me.

relationship	generations	segment threshold in cM						
		0	5	7	10	20	30	50
1st cousins	2.0	100	100	100	100	100	100	99
1st cous 1 rem	2.5	100	100	100	100	100	99	79
2nd cousins	3.0	100	100	100	100	98	87	42
2nd cous 1 rem	3.5	100	99	99	98	84	58	17
3rd cousins	4.0	98	94	91	85	57	30	6.2
3rd cous 1 rem	4.5	90	78	72	62	32	14	2.1
4th cousins	5.0	72	55	48	39	16	5.9	0.72
4th cous 1 rem	5.5	51	35	29	22	7.6	2.5	0.24
5th cousins	6.0	32	20	16	11	3.5	1.0	0.08

One obvious question is to ask how likely is a cousin of a certain degree to match me, for a given segment length? The table above gives the simulation's answer for a few chosen segment lengths. Taking the 50 cM threshold as an example, 99% of first cousins will match you at 50 cM or better, but only 42% of second cousins, and 6% of third cousins.

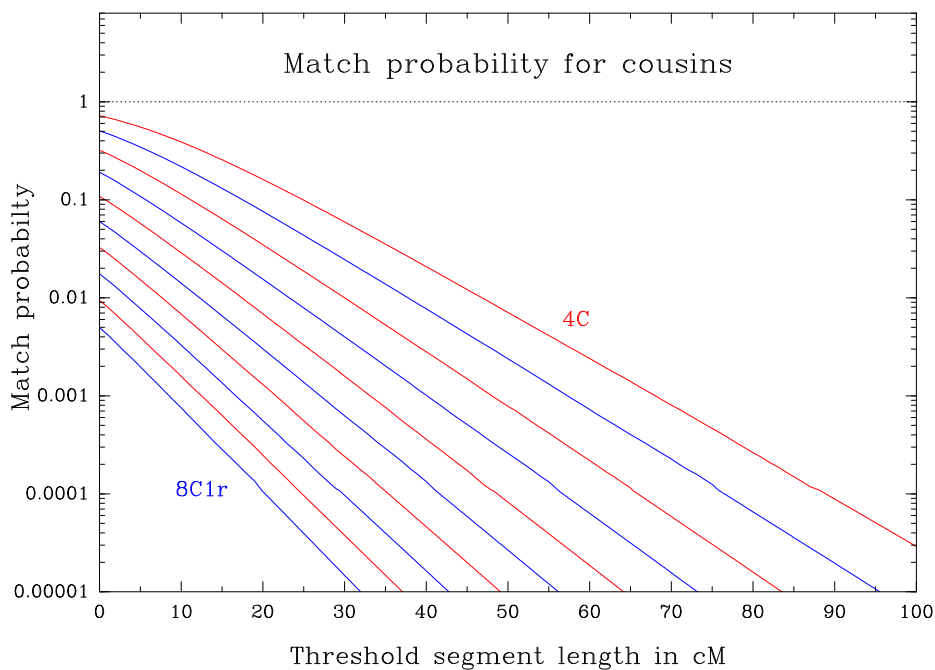
Once we get to third cousins or beyond, we see that there is a fair chance that a cousin will share no DNA at all, however small the threshold is made. Only 32% of 5th cousins have any DNA shared with you, only 11% have a match that's longer than 10 cM. Of course, there is a fact working in the opposite direction; you have very many more 5th cousins than 2nd cousins, so you will see a number of matches way out on the tail.

We can show these threshold probabilities in a graph.



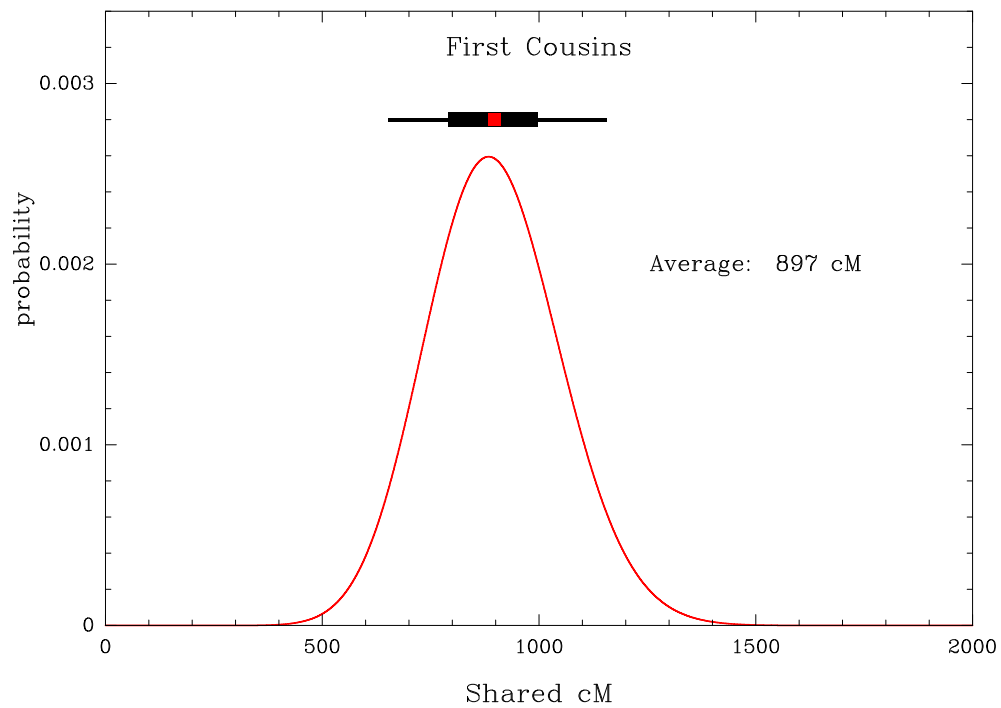
First cousins are almost certain to match you (probably in many places) with segments of 50 cM, and quite a few may match at 100 cM. Third cousins are very unlikely to match at 70 cM or more.

For distant cousins (4th cousins, out to 8th cousins once removed), I've used a logarithmic scale so that very small probabilities can be shown. Because the number of very distant cousins is so large, there can be a few who match far out on the probability tail.

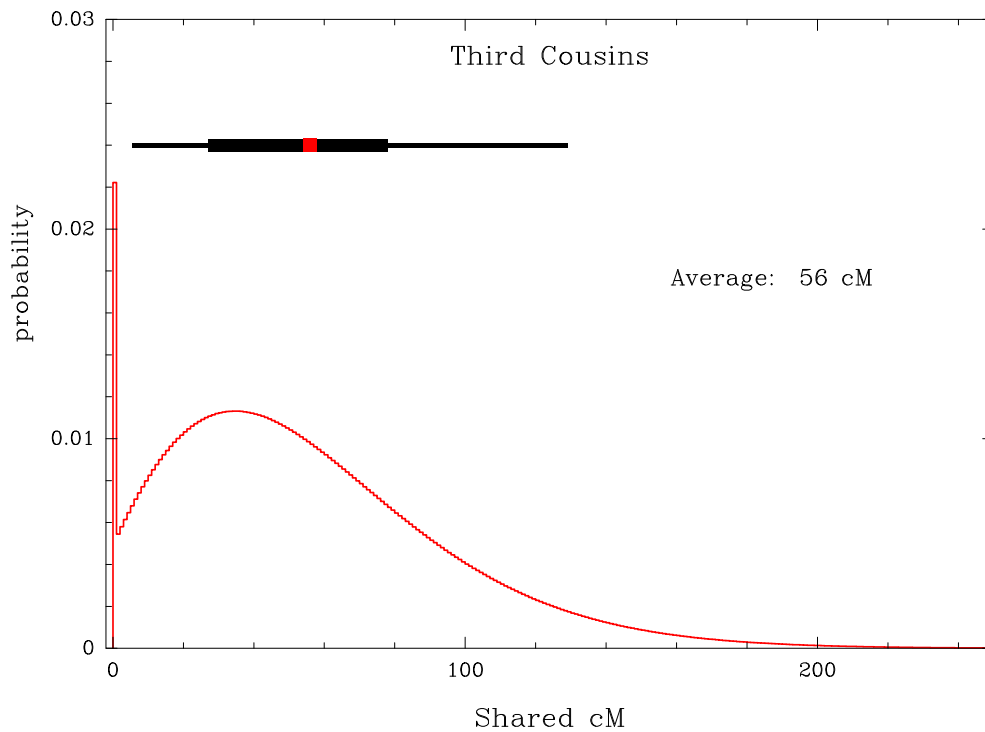


6 First cousins

As well as calculating averages, we can also use the simulation to give us widths and probability distributions for the quantities we are interested in.

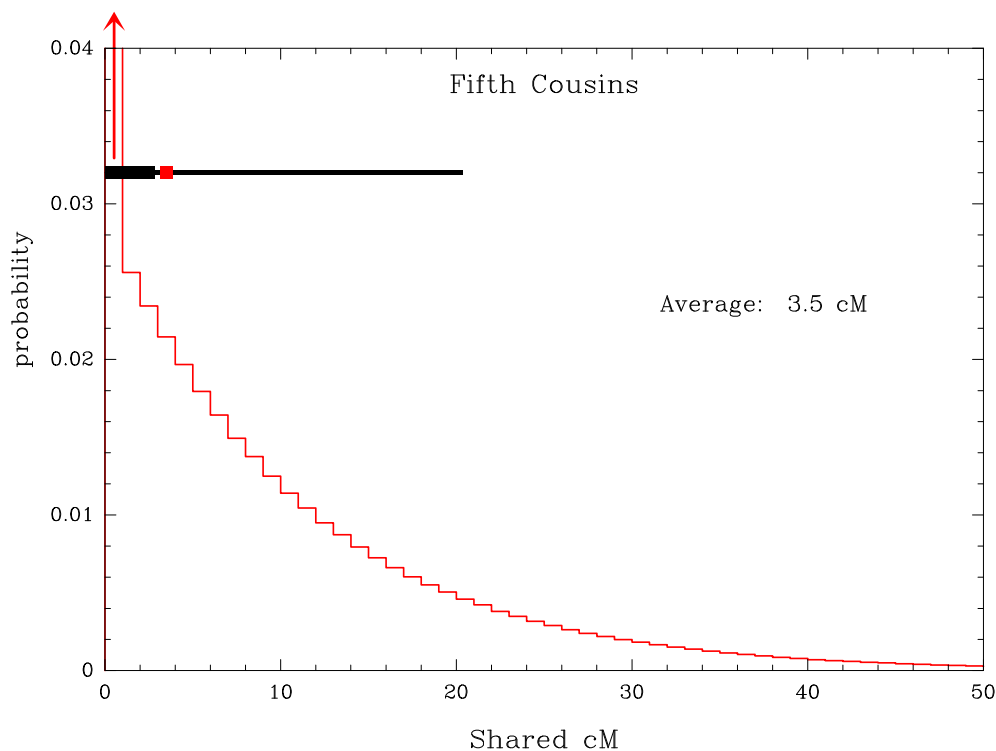


7 Third cousins



For close cousins shared DNA follows a symmetric bell curve, but as we move to more remote relationships the probability becomes more and more lopsided. With third cousins there is already a 2% chance that a cousin will completely fail to match, and the scatter of the cousins who do match is very wide.

8 Fifth cousins



As we move out to more distant relationships, we lose all resemblance to the bell-shaped curve. We have a large number of cousins who share zero DNA, but the few who do share, can still share quite large amounts.

For 5th cousins, the simulation says that 68% share no DNA, 32% share some DNA. If we average over all 5th cousins, the average shared DNA is 3.5 cM. But a more useful average might be the average when we forget all about the undetectable cousins who don't match us at all, and just consider the 32% who do match. For this sub-sample the average amount of shared DNA is a more respectable 10.9 cM.